基于电子病历的实体识别和知识图谱构建的研究*

黄梦醒[a, b]，李梦龙[a, b]，韩惠蕊[a, b]

(海南大学 a. 南海海洋资源利用国家重点实验室; b. 信息科学技术学院，海口 570228)

摘　要：针对中文电子病历中命名实体识别和实体关系抽取研究方法中存在的问题，提出了一种基于双向长短时记忆网络(bidirectional long short term memory)与CRF(conditional random field)结合的实体识别和实体关系抽取方法。该方法首先使用词嵌入技术将文本转换为数值向量，作为神经网络BiLSTM的输入，再结合CRF链式结构进行序列标注，输出最大概率序列，并对识别结果知识图谱化。实验证明，该方法对中文电子病历进行实体识别和实体关系抽取时的准确率、召回率、F值有明显的提升。实验结果满足临床中系统应用需求，对帮助研究构建临床决策支持系统、个性化医疗推荐服务有引导作用。

关键词：实体识别；实体关系；长短时记忆网络；知识图谱

中图分类号：TP391.1　　doi: 10.3969/j.issn.1001-3695.2018.07.0414

Research on entity recognition and knowledge graph construction based on electronic medical records

Huang Mengxing[a, b], Li Menglong[a, b], Han Huirui[a, b]

(*a. State Key Laboratory of Marine Resource Utilization in South China Sea, b. College of Information Science & Technology Hainan University, Haikou* 570228, *China*)

Abstract: Aiming at the problems in the research methods of named entity recognition and entity relationship extraction in Chinese electronic medical records) , this paper proposed an entity identification and entity relationship based on bidirectional long short term memory and conditional random field (CRF) . The method first used word embedding technology to convert text into numerical vector, as the input of neural network BiLSTM, combined with CRF chain structure for sequence labeling, output the maximum probability sequence, and mapping the recognition result knowledge graph by using the database tool Neo4j. Experiments show that the method can significantly improve the accuracy, recall rate and F value of entity identification and entity relationship extraction in Chinese electronic medical records. The experimental results meet the needs of clinical system applications, and have a guiding role in helping to study and construct clinical decision support systems and personalized medical recommendation services.

Key words: entity recognition; entity relation; BiLSTM; knowledge graph

0　引言

电子病历（electronic medical record,EMR）是指医务人员在对患者医疗的过程中，使用医疗机构信息系统生成的文字符号、、图表、图形、数据等数字化电子信息，还有存储、管理、传输和重现医疗记录的作用。同时，EMR也是一种非常宝贵的知识资源，其中包含了大量的、准确的详细的患者的医疗信息。通过对电子病历完成知识提取任务，获得患者详细的医疗信息，一方面可以帮助医学研究者构建临床决策支持系统；另一方面还可以作为辅助信息，帮助医生解决知识上的局限性问题，从而减少个人的医疗失误问题。此外，对电子病历完成知识提取，在未来的工作中为患者或用户提供高效便捷的个性化医疗推荐服务，做足准备工作。

知识图谱的构建是当前各研究领域的又一大热点。知识图谱本质是语义网络技术，由Google于2012年提出，主要目的在于提高互联网的搜索效率。将真实世界中事物与事物之间的联系转换为知识图谱中实体与实体之间的关系来描述。现阶段，国内医疗领域中基于知识图谱的疾病预测研究工作才刚起步，

收稿日期：2018-07-24；**修回日期：**2018-09-07　　**基金项目：**国家自然科学基金资助项目（61462022）；国家科技支撑计划项目（2015BAH55F04）；海南省重大科技计划项目（ZDKJ2016015）；海南省自然科学基金资助项目（617062）；海南省产学研一体化专项资助项目（cxy20150025）

作者简介：黄梦醒（1973-），男，河南信阳人，教授，博导，博士，主要研究方向为大数据与智慧服务（huangmx09@hainu.edu.cn）；李梦龙（1992-），男，河南新乡人，硕士研究生，主要研究方向为医疗大数据与自然语言处理；韩惠蕊（1990-），女，海南海口人，博士研究生，主要研究方向为推荐系统、机器学习和医学信息学.

所以构建基于知识图谱的医疗知识系统对于智慧医疗的发展具有一定的辅助意义。

1　相关工作

20 世纪末，医疗信息化在国际上的发展，已经达到了一定的成熟阶段，具有大规模的语料库和研究方法，还建立了一体化医学语言系统（unified medical language system，UMLS）。在医学领域的研究中，自然语言处理（natural language processing，NLP）中的实体识别（entity recognition，NER）和实体关系抽取（relation extraction，RE）一直是热点与难点。

在信息提取阶段，实体识别的主要任务是从电子病历中找到当前知识架构基础上已经存在的概念词语，其中包括疾病，病症，药物，检测，治疗等；实体关系抽取的主要任务是发现并建立两个实体之间的关系，包括疾病和病症之间的关系，疾病和药物之间的关系等。这两个阶段也使得未来构建个性化医疗健康服务系统有了一个很好的准备工作。

关于实体识别和实体关系提取的研究，广泛应用的方法可分为三类：基于词典的方法；基于过则的方法；基于机器学习方法。龙光宇等人[错误!未找到引用源。]采用条件随机场（CRF）与基于词典相结合的方法对疾病进行实体识别，王宁等人[错误!未找到引用源。]通过使用手动构建的规则来识别金融领域中的公司名称。基于词典与规则的方法太过于依赖词典，规则等人工预料库的构建，泛化能力弱，可移植性差。基于机器学习的实体识别方法通常可分为两类，一类是基于分类的方法，另一类是基于将实体识别问题转化为序列的整体标注问题，即同时对一段话中多个词进行标记，最后选择联合概率最大的标注序列，有较强的扩展性和适应性。例如，传统的序列标注一般使用“BIO”标注方法，在实体识别过程中加入一个实体类别标签“C”，标签形式为“BIO+C”。其中，“B”表示一个实体的开始，“I”表示实体的继续，“O”代表不属于已定义的任何一种实体，“C”为实力类别标签。在语料库的构建过程中，需要统一规范。曲春燕等人[27]结合中文电子的语言结构特点，结合原有的电子病历标注规范，制定了较为详细《中文电子病历命名实体和实体关系标注规范》。为自然语言处理在中文电子病历领域的研究创建了很好的基础。Li 等人[错误!未找到引用源。]对比了 CRF 和支持向量（support vector machine，SVM）在电子病历实体识别中的性能，实验结果表明 CRF 有较好的性能。Lample 等人[错误!未找到引用源。]提出了 LSTM+CRF 模型，并证明该模型性能超过了 CRF 模型性能，最大的优点在于无需特征工程，使用词向量就可以达到很好的效果。在医疗领域，关于实体关系抽取研究，Uzunerd 等人[错误!未找到引用源。]率先定义了医疗实体关系。在自然语言处理的其他领域也有应用,Socher 等人[错误!未找到引用源。]提出使用循环神经网络（recurrent neural network ,RNN）处理实体关系抽取问题。

在实体识别和实体关系抽取的基础上，通过采用率最高的图数据库 Neo4j 对电子病历中的疾病，病症，以及它们之间的关系，以图形化的方式显示出来，更能增强医疗服务的便捷性和医疗知识的可理解性。

本文通过总结医疗研究领域中应用广泛的实体识别方法和实体关系抽取方法，提出一个新的框架模型结构，并在此基础上使用 Neo4j 图数据库对知识进行管理和可视化，对推动智慧医疗发展有一定的指导作用。本文将任务分为两个阶段：

a)知识提取阶段。在该阶段，以实体识别和实体关系抽取为主要任务。在已构建语料库的基础上，使用自然语言技术对电子病历进行自动识别，采用实体识别和实体关系抽取的机器学习方法对其进行分析，抽取，构建医疗实体之间的关系。

b)知识存储阶段。该阶段的主要任务是以图结构存储知识并通过 Neo4j 实现可视化阶段，通过知识图谱的形式，将疾病实体、病症实体以及药物实体及其相互之间的关系关联起来。模型大致框架如图 1 所示。

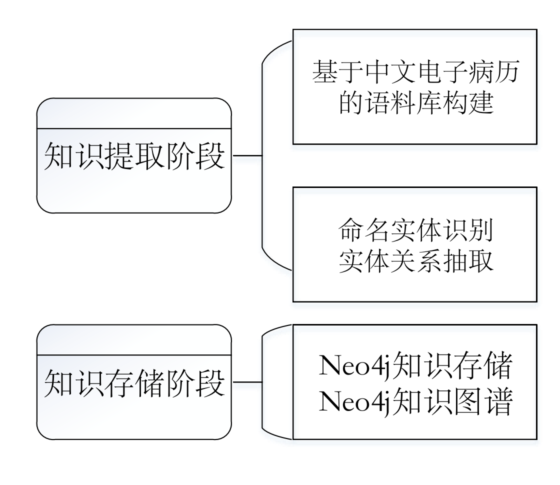


图 1　总体模型框架

2　语料库构建

通过对中文电子病历的文本特点进行分析，在文献[7,8]的基础上，制定给了相应的标注规范其结构如图 2 所示。本文构建语料库的数据来源于海口市中医院提供的 2 300 份中文电子病历，共包含 15 个大小不同的科室。在构建语料库之前，先对数据进行去敏处理，然后从不同科室中随机挑选出一定量的电子病历进行数据标注。已完成标注的中文电子病历共 500 份。下图统计分析了公开数据集新闻语料和中文电子病历，根据显示结果发现电子病历实体分布密集程度远高于新闻语料。

3　知识提取模型设计

基于词典和规则的方法，虽然在实体识别中的实验有较好效果，但是考虑到构建专业词典和规则难度过大，且实验方法泛化能力弱。本文参考之前实体识别模型思想，并对存在的问题进行分析统计，然后采用了一种全新的模型思想，并以中文电子病历语料为训练数据进行实验探究。知识提取是指从不同数据源中挖掘目标知识，本文主要知识来源是从中文电子病历中提取的患者信息，通过使用实体识别方法和实体关系抽取方法将文本电子病历中的信息提取出来，并进行分析应用，知识提取模型框架如图 3 所示。

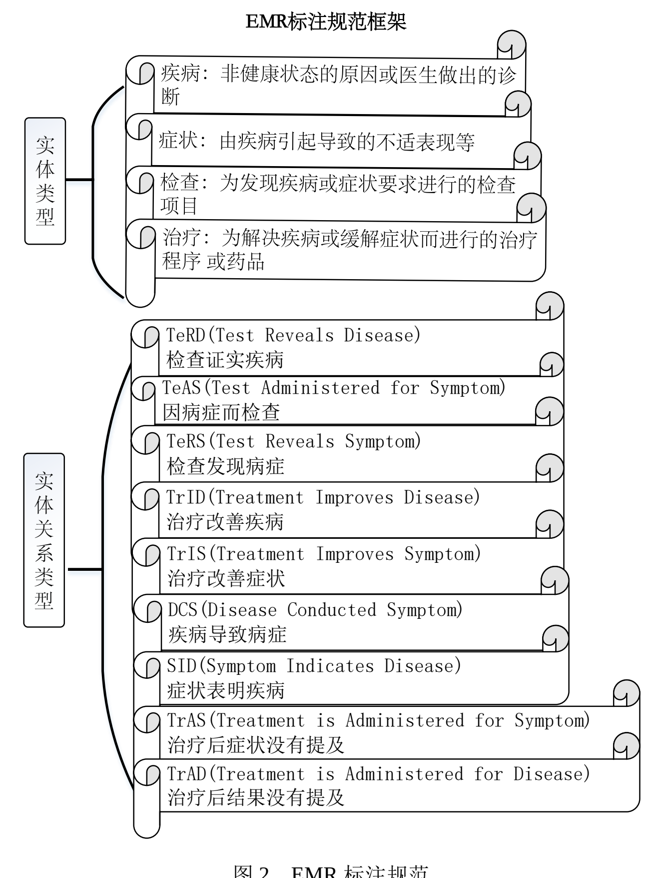


图 2　EMR 标注规范

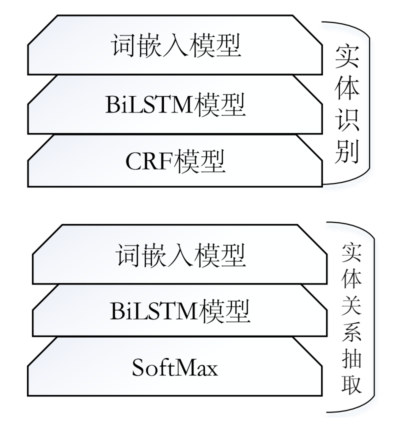


图 3　知识提取框架

3.1　实体识别模型

实体识别模型分为三层,第一层是词嵌入模型，作用是将文本转换为数词向量,将词向量输入到第二层；第二层是 BiLSTM 模型，作用是以词向量为输入，自动提取文本特征，将文本特征作为 CRF 线性层的输入；第三层是 CRF 模型，作用是对 BiLSTM 提取出来的文本特征进行序列标注，并从句子的整体层面考虑，达到全局最优序列。

3.1.1 词嵌入模型

词嵌入（word embedding，WE）一项非常重要且应用非常广泛的技术，可以将文本和词语转换为计算机可以识别的数值向量。主要分为两类：一类是词袋模型（bag of words ,BOW），一类是分布式表示（distributed representation）。BOW 的代表是 one hot 编码，虽然起源很早，但由于不能很好的保存语义信息，不考虑使用。本文采用的是分布式表示方法中的 word2vec，是 Google 在 2013 年提出的一个开源 NLP 工具，特点是将所有词向量化，可以定量的度量词与词之间的关系，进一步挖掘词之间更深的含义。该方法有两种训练模型，CBOW（continuous bag-of-words model）和 skip-gram（continuous skip-gram model）。CBOW 的思想是将一个词的上下文作为输入，词本身作为输出，通过上下文推测词的含义；skip-gram 则是将词本身作为输入，词的上下文作为输出，本文采用后者，训练并得到相应的词向量。

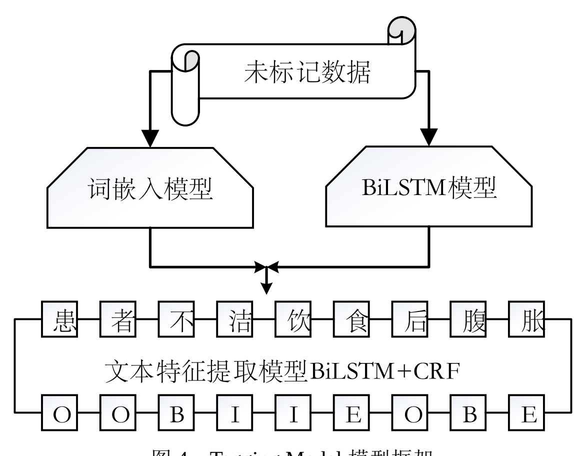


图 4　Tagging Model 模型框架

研究表明，word2vec 将词语从文本转换为词向量，虽然保留了句子的语义特征，但是并不能完整的保存句子本身的一些句法结构上的特征。为了更好的保留语义和结构特征，本文尝试在文本预处理阶段加入一个 BiLSTM 预测下一时刻的字符，进一步表示文本的语义和句法结构。将 Word2Vec 与 BiLSTM 的隐藏层向量联合作为文本特征提取模型的输入，将其命名为 Tagging Model，其结构如图 4 所示。

3.1.2 文本特征提取模型 BiLSTM

在众多实体识别方法模型中，由于基于机器学习方法有较强的扩展性和适用性，得以广泛使用。但还是存在诸多问题，本文针对其中的一些问题，并作出改进，加入到新的模型当中。

在传统神经网络中，存在问题有：a）不同层次之见的神经元全连接，相同层次之间无连接;b）不能捕捉前后词语标注后的结果。采用隐藏层之间有连接的 RNN 解决以上问题，但在实际应用中，由于梯度弥散问题，通常假设当前状态只与之前邻近的节点状态有关，降低模型的复杂程度。在 RNN 中存在长期依赖问题，是指经过许多阶段传播后的梯度倾向于消失和爆炸。为解决 RNN 中的长期依赖问题，最有效的方法是 LSTM（long short term memory）结构，由 Hochreiter 等人[28]提出，后被 Alex Graves 改进，得到了广泛的使用。

为了能够有效地利用上下文的信息，将标准的 RNN 单向时序处理方式拓展为双向 LSTM（BiLSTM）网络，模型内包含两个方向的网络结构：方向 1 是从左到右 $(\overrightarrow{h_1},\overrightarrow{h_2},\overrightarrow{h_3},\overrightarrow{h_4})$（顺序）传播；方向 2 是从右向左 $(\overleftarrow{h_1},\overleftarrow{h_2},\overleftarrow{h_3},\overleftarrow{h_4})$（逆序）传播，采用向量拼接的方式将其组合，通过连接的线性层将其映射为 k 维，k 是训练集内的标签种类数，从而得到提取的句子特征，记作

(c_1,c_2,c_3,c_4)。其结构如图 5 所示。

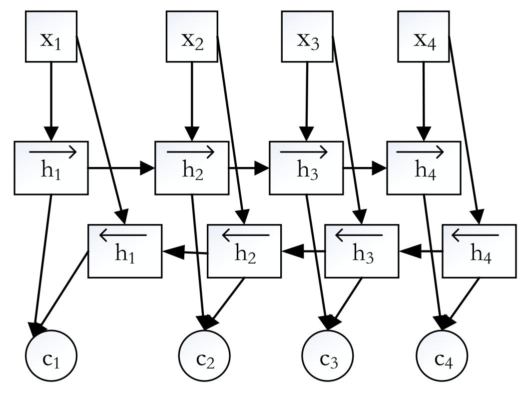


图 5　BiLSTM 模型框架

3.1.3 序列标注模型 CRF

虽然 LSTM 可以完成序列标注任务，但存在标注偏置问题。使用 CRF 模型完成序列标注任务，可以获取全局最优输出序列，效果优于单个 LSTM。使用 CRF 链式结构，如图 6 所示。将 BiLSTM 提取的文本特征 $P(Y\mid C)$ 进行序列标注，标注方法采用"BIEOS"，单个字母分别代表实体开始、中间、结束，无关实体，单个字符实体，如图 3.2 中标注所示。令 $C=(c_1,c_2,c_3,c_4)$ 和 $Y=(y_1,y_2,y_3,y_4)$ 分别作为 CRF 链式结构的观察序列和状态序列，即输入和输出。$P(Y\mid C)$ 是在 C 序列的条件下 Y 序列的条件概率分布，计算过程如下：

$$\mathrm{P(Y|C)}=\frac{1}{Z(C)}\exp\left(\sum_i\sum_j\lambda_j f_j(y_{i-1},y_i,\mathrm{C,i})+\sum_i\sum_j u_j s_j(y_i,\mathrm{C,i})\right) \quad (1)$$

$$\mathrm{Z(C)}=\sum_y\exp\left(\sum_i\sum_j\lambda_j f_j(y_{i-1},y_i,\mathrm{C,i})+\sum_i\sum_j u_j s_j(y_i,\mathrm{C,i})\right) \quad (2)$$

$$y^*=argmaxP(Y\mid C) \quad (3)$$

其中：概率转移函数 $\mathrm{f_j(y_{i-1},y_i,C,i)}$ 表示序列 C 在 y_{i-1} 到 y_i 之间的转移概率。状态函数 $\mathrm{s_j(y_i,C,i)}$ 表示序列 X 在第 i 个位置的标记为 y_i 的概率， $\mathrm{Z(C)}$ 是归一化项。λ_j 和 u_j 分别对应相应函数的权重。$\mathrm{y^*}$ 表示最有可能的标注结果序列，即最大条件概率。

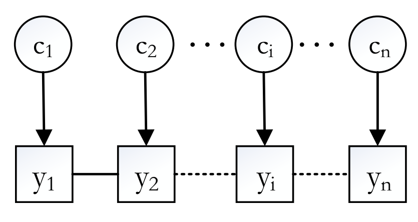


图 6　CRF 链式结构

3.2　实体关系抽取模型

电子病历命名实体关系抽取任务的主要目的是研究识别两个医疗实体间的预定义的关系，例如疾病、症状、检查和治疗这几类实体间的关系。当前国内外研究中普遍使用的是将实体关系抽取作为一个独立的多分类问题，或将实体关系与实体识别串联组合。存在的问题是忽略了两个任务之间的关联，例如当实体识别错误时，会进一步影响下一个实体关系抽取错误，从而扩大整体的错误率。

传统的实体关系抽取方法，首先是先进行实体识别，然后对识别的实体进行分类完成关系的抽取，这种方法叫做流水线方法。虽然模型的灵活性较高，但不足是实验之前需要具有专业领域知识的人对数据进行数据标注处理，会消耗大量时间和人力。而且由于训练集被大量标注，包含了一定的先验知识，会影响模型的识别能力。

本文在实体识别模型的基础上对实体关系抽取进行研究，将两者尝试作为联合任务，提出了新的标注方法，将文本的位置标签使用 one hot 编码，作为辅助信息输入到模型中并做出合适的重构来更加符合任务的完成。结合实体识别设计一个联合模型 Entity Relation Model（ERM），采用"BIEOS"标注。将原标注方法添加一组预定关系，转换为三元组，如（实体信息，实体关系，实体在关系中的位置）B-TeRD-1、E-TeRD-2。本文只考虑一个实体只属于一个三元组的情况。ERM 模型与实体模型类似，将实体识别模型中的 CRF 层更换为 Softmax 层。模型结构如图 7 所示。

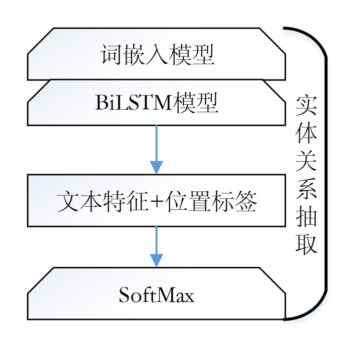


图 7　ERM 模型框架

在实体关系抽取模型中，共有四层，第一层词嵌入模型，将文本内容转化为词向量；第二层 BiLSTM 模型，自动提取文本特征；第三层是附加位置标签的文本特征，通过使用 one hot 编码将向量化的文本特征与实体位置标签联合，构成一个三元组格式；第四层是采用处理多分类问题的 Softmax 函数层，将第三层的关系分类转换为最大概率问题。ERM 模型如图 8 所示。

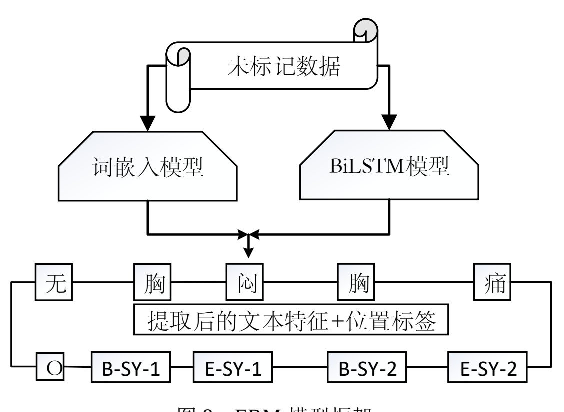


图 8　ERM 模型框架

3.3 Neo4j 知识存储模型

本文采用自底向上构建方式构造知识图谱。在众多数据库系统中，Neo4j 具有高性能，设计灵活性，开发敏捷性等优势，用户可以使用 Cypher 语言来操作数据。通过实体识别模型和实体关系抽取模型，将输出结果使用 Python 转换为 SPO 三元组（Subject, Predicate, Object）,如表 1 所示,整个知识图谱可以看作三元组集合。

表 1　SPO 三元组

Subject	Predicate	Object
乙型肝炎	症状	腹痛
乙型肝炎	症状	皮肤巩膜黄染
拉米夫定	治疗	乙型肝炎

通过对 150 多种疾病相关的三元组分析统计，发现某一特定症状可以由多种疾病引起，因此在将病症实体导入到 Neo4j 数据库时，我们需要对每一个病症节点设置唯一性约束。在衡量单个病症对于疾病的影响因子上， Rotmensch 等人错误!未找到引用源。提出了一种基于朴素贝叶斯和知识图谱的病症权重因子 IMPT 计算方法，即

$$\mathrm{IMPT}=\log\left(\mathrm{P}\left(x_i=1|y_j=1\right)\right)-\log\left(\mathrm{P}\left(x_i=1|y_j=0\right)\right) \tag{4}$$

其中:IMPT 表示单个病症对疾病的权重因子, x_i 表示 $id=i$ 的病症实体， y_j 表示 $id=j$ 的疾病实体,值“1”“0”表示疾病和病症的有无。IMPT 值越大,就表示知识图谱中连接对应疾病实体与病症的权重越大。

4 实验分析

4.1 实体识别实验结果分析

在训练模型过程中，结合 Bootstrapping 方法调整训练过程，进一步扩展数据集。训练过程如下所示：

a)根据已有标注数据，训练好初始的 Tagging Model；

b)获取未标注语料，输入到训练好的 Tagging Model,得到一个分类标签以及概率，若概率大于阈值，则将词语与标注结果加入可靠集中；

c)当可靠集数量 N=500 时，将现有的可靠集与原标注数据集合并为训练集，重新训练 Tagging Model，并清空可靠集，重复步骤 b)。

实验过程中，将海口市提供的 2300 份电子病历作为实验语料。采用交叉验证方法，已标注的 500 份语料中，随机抽取 100 份作为训练语料，400 份作为测试语料。通过对比不同参数的模型结构，设置词嵌入向量维数采用 256 维，隐藏层数 4（每个方向两层），优化算法使用自适应时刻估计方法（adaptive moment estimation, Adam），损失函数使用交叉熵损失函数，学习率 0.001，dropout 0.3。实体识别针对图 2 中的 4 种实体类型进行实验，准确率、召回率和 F 值作为评估标准。实验结果如图 9 所示。

从图 9 可以发现，相对于单独使用 BiLSTM 对实体识别，BiLSTM 与 CRF 的联合对识别结果有一定的提高。而使用词向量作为特征的 CRF 模型效果低于手工提取特征的 CRF 模型。B-tagging Model 是在训练过程中使用 Bootstrapping 方法扩展数据集，提升了模型的识别效果。同时观察发现准确率和召回率都有提高，说明模型的泛化能力得到进一步的提升，增强模型的适用性。图 10(a)(b)分别表示实体识别模型 B-Tagging Model 对图 2 中 4 种实体类型的准确率和召回率评估，结果表明实体类型检查与治疗的准确率都很高，原因可能是因为这两类实体有特殊的结构和语法特点。对于疾病和症状可能是因为在电子病历中的位置都很相似导致分类失误。

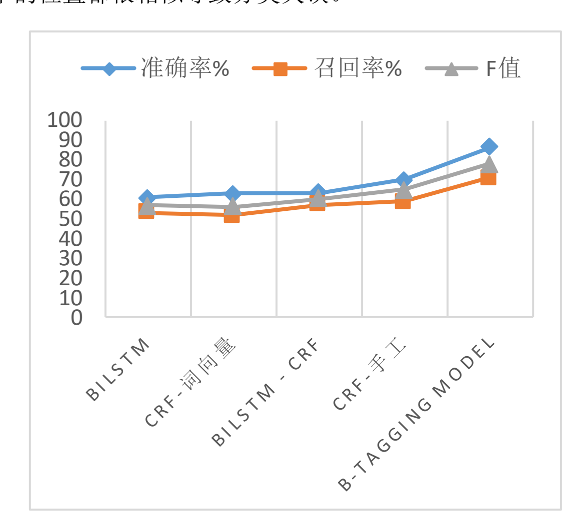
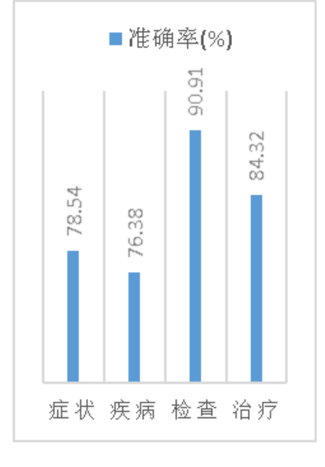
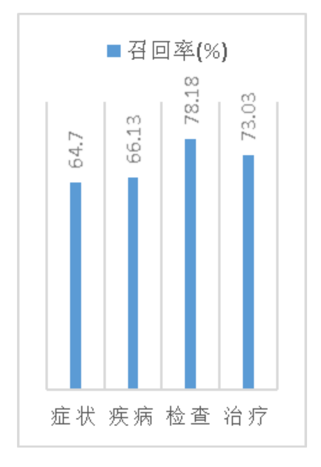


图 9　五种实体识别方法实验结果



(a) 准确率测试



(b)召回率测试

图 10　B-Tagging Model 对不同实体类型识别结果

4.2 实体关系抽取实验结果分析

对于实体关系抽取模型的实验，与上一节实体识别模型相似，实验参数设置相同。实验结果如图 11 所示。

由图 11 可以发现，相比于传统的 CRF 模型，ERM 模型虽然准确率有所提高，但是召回率和 F 值都低于 CRF 模型。相比于 ERM-T（添加位置标签后），前两种方法都有所不足。ERM-T 的是 3 个评估标准都高于 CRF 和 ERM。图 12 中 a, b 图是实体关系抽取模型 ERM-T 对于图 2 中 9 中实体关系识别结果的准

确率和召回率评估。结果表明 TrID 和 DCS 效果最好，TrAD 和 TrAS 最差，究其原因可能是因为相同的位置标签内，可以较好的识别出实体关系，若间隔较大，则效果明显降低。

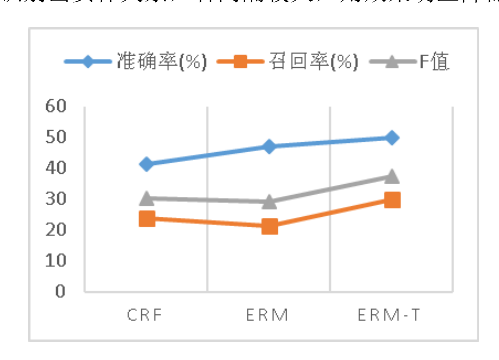
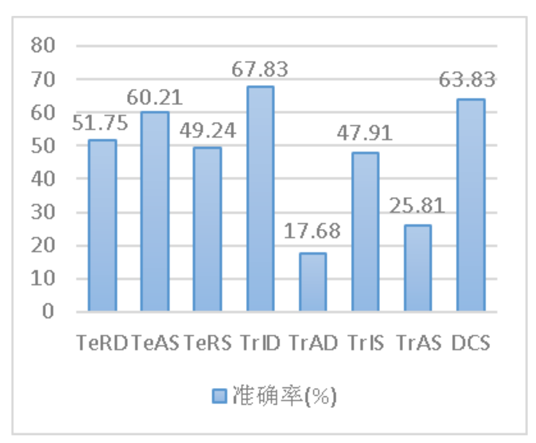
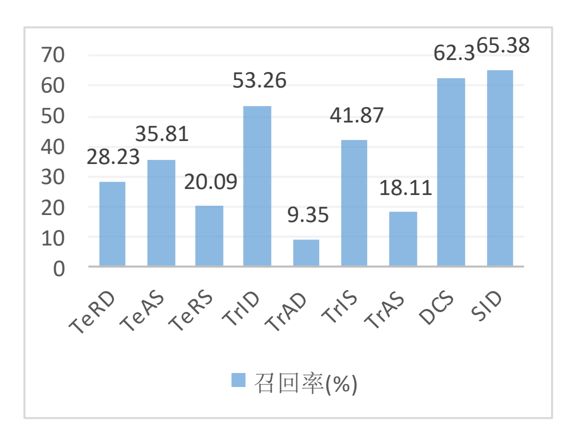


图 11　三种实体关系抽取模型实验结果



(a)准确率测试



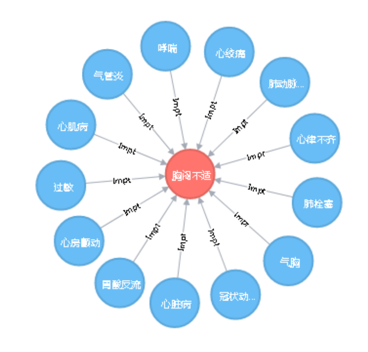
(b)召回率测试

图 12　ERM-T 对实体关系抽取结果

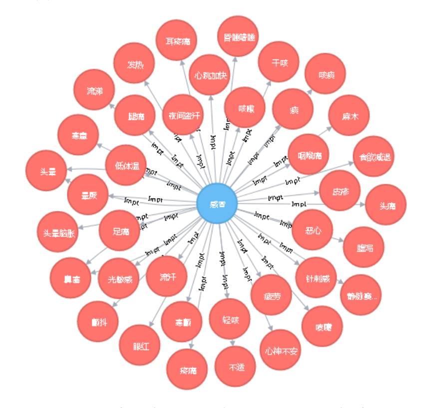
4.3　知识图谱可视化结果分析

通过将三元组结构化数据，使用 Java 将数据传送到 Neo4j 本地数据库中，并构建好知识图谱。如图 13 所示。

在 Neo4j 图存储数据库管理系统中，患者，医生或者用户可以通过输入症状或疾病，通过分析，与构建好的医疗知识图谱进行匹配，根据病症实体或疾病实体与药物实体之间的关系，推荐相关疾病知识和药物，如何预防疾病和饮食习惯等相关知识给用户。通过构建医疗知识图谱系统，不仅患者可以查询自己可能患有的疾病，还可以帮助医务人员查阅相关疾病信息，达到辅助医疗的作用。



(a)病症“胸闷不适”与各种疾病的联系



(b)疾病“感冒”与各种症状的联系

图 13　知识图谱可视化结果

图 13 中的(a)是在 Neo4j 数据库中使用 match 语句搜索病症“胸闷不适”后的结果，(b)是搜索疾病“感冒”后的结果。蓝色节点表示疾病实体，红色节点表示病症实体，两者之间的联系是 IMPT 权重因子。通过可视化操作，我们能够很清晰的看到单一疾病与多种病症以及单一病症与多种疾病之间的联系。

5　结束语

由于中文电子病历的独特文本特点和缺少大规模语料库，且没有广泛的统一标注规范，致使研究问题重重。本文基于在当前研究领域广泛使用的各种模型方法，基于词典，基于规则和基于机器学习的命名实体识别方法，结合词嵌入技术和神经网络 BiLSTM 以及 CRF 模型，构造一个新的模型用于命名实体识别，研究证明该模型有良好的表现。在实体识别基础上，分析传统实体关系抽取方法存在的分离问题，联合实体识别，构建新的模型方法，提升了识别效果。在训练过程中，还使用 Bootstrapping 方法，扩展训练语料，增量模型的有效性。

虽然模型方法取得良好效果，但是模型方法还有很多不足之处，有待进一步完善。在实体关系抽取模型中，不同类别实体间隔较远，BiLSTM 不能有效的发现两者之间的关系，可以尝试使用注意力机制或者根据文本内容建立分布度加权帮助模型学习不同类型实体间的关系。还可以尝试利用多分类思想以及句法树的思想改进模型，发现实体之间的关联。

本文实现的知识图谱功能尚有不足，有待进一步拓展。可以尝试将医学知识扩展到语料库，识别新的实体类型，构建新的实体关系，更加完善医疗体系知识图谱的应用性。

参考文献：

[1] 龙光宇，徐云. CRF 与词典相结合的疾病命名实体识别 [J]. 微型机与应用, 2017, 36 (21): 51-53. (Long Guangyu, Xu Yun. Combining CRF and dictionary based disease named entity recognition [J]. Information Technology and Network Security, 2017, 36 (21): 51-53)

[2] Wang Ning, Ge Ruifang, Yuan Chunfa, *et al.* Company name identification in Chinese financial domain [J]. Journal of Chinese Information Processing, 2002, 16 (2): 1-6.

[3] Li Dingcheng, Karin K S, Guergana S. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts [C]// Proc of Workshop on Current Trends in Biomedical Natural Language Processing. Stroudsburg: ACL, 2008: 94-95.

[4] Lample G, Ballesteros M, Subramanian S, *et al.* Neural Architectures for Named Entity Recognition [J]. Computation and Language, 2016: 260-270.

[5] Uzuner O, Mailoa J, Ryan R, *et al.* Semantic relations for problem-oriented medical records [J]. Artificial Intelligence in Medicine, 2010, 50 (2): 63-73.

[6] Socher R, Huval B, Manning C D, *et al.* Semantic compositionality through recursive matrix-vector spaces [C]// Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 1201-1211.

[7] 杨锦锋，关毅，何彬，等. 中文电子病历命名实体和实体关系语料库构建 [J]，软件学报, 2016, 27 (11): 2725-2746. (Yang Jinfeng, Guan Yi, He Bin, *et al.* [J]. Journal of Software, 2016, 27 (11): 2725-2746.)

[8] 赵芳芳. 面向中文电子病历的词性标注技术研究 [D]. 哈尔滨：哈尔滨工业大学, 2014. (Zhao Fangfang. Research on part-of-speech tagging for Chinese electronic medical records [D]. Harbin Institute of Technology, Harbin, 2014)

[9] Rotmensch M, Halpern Y, Tlimat A, *et al.* Learning a health knowledge-graph from electronic medical records [J]. Scientific Reports, 2017, 7 (1): 1-11.

[10] 杨锦峰，于秋滨，关毅，等. 电子病历命名实体识别和实体关系抽取研究综述 [J]. 自动化学报，2014, 40 (8): 1537-1562. (Yang Jinfeng, Yu Qiubin, Guan Yi, *et al.* An overview of research on electronic medical record oriented named entity recognition and entity relation extraction [J]. Acta Automatica Sinica, 2014, 40 (8): 1537-1562)

[11] 张立邦. 基于半监督学习的中文电子病历分词和名实体挖掘 [D]. 哈尔滨：哈尔滨工业大学，2014. (Zhang Libang. Word segmentation and named entity mining based on semi supervised learning for Chinese EMR [D]. Harbin Institute of Technology, Harbin, 2014)

[12] 李航. 统计学习方法 [M]. 北京：清华大学出版社，2012. (Li Hang. Statistical learning methods [M]. Beijing: Tsinghua University, 2012)

[13] 胡章荣，王朝斌. 基于词典的中文分词算法及其性能评估 [J]. 电子技术与软件工程，2015 (15): 102-106. (Hu Zhangrong, Wang Chaobin. Chinese word segmentation algorithm based dictionary and its performance evaluation [J]. Electronic Technology and Software Engineering, 2015 (15): 102-106.)

[14] 贾李蓉，刘静，于彤，等. 中医药知识图谱构建 [J]. 医学信息学杂志，2015, 36 (8): 51-53. (Jia irong, Liu Jing, Yu Tong, *et al.* Construction of traditional Chinese medicine knowledge graph [J]. Journal of Medical Intelligence, 2015, 36 (8): 51-53.)

[15] 刘峤，李杨，段宏，等. 知识图谱构建技术综述 [J]. 计算机研究与发展，2016, 53 (3): 582-600. (Liu Qiao, Li Yang, Duan Hong, *et al.* Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53 (3): 582-600.)

[16] 栗伟，赵大哲，李博，等. CRF 与规则相结合的医学病历实体识别 [J]. 计算机应用研究, 2015, 32 (4): 1082-1086. (Li Wei, Zhao Dazhe, Li Bo, *et al.* Combining CRF and rule based medical named entity recognition [J]. Application Research of Computers, 2015, 32 (4): 1082-1086.)

[17] 李丽双，郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别 [J]. 中文信息学报, 2018, 32 (1): 116-122. (Li Lishuang, Guo Yuankai. Biomedical named entity recognition with CNN-BLSTM-CRF [J]. Journal of Chinese Information Processing, 2018, 32 (1): 116-122.)

[18] 李晓静，林海伦，贾岩涛，等. 融合结构与内容的在线百科实体标注方法 [J]. 计算机科学与探索，2015, 9 (10): 1238-1246. (Li Xiaojing, Lin Hailun, Jia Yantao, *et al.* Online encyclopedia entities tagging method based on page structure and content [J]. Journal of Frontiers of Computer Science and Technology, 2015, 9 (10): 1238-1246.)

[19] 戴雪，蒋志鹏，关毅. 基于中文电子病历的跨科室组块分析 [J]. 计算机应用研究, 2017, 34 (7): 2084-2087. (Dai Xue, Jiang Zhipeng, Guan Yi. Cross-department chunking based on Chinese electronic medical record [J]. Application Research of Computers, 2017, 34 (7): 2084-2087.)

[20] 秦长江，侯汉清. 知识图谱——信息管理与知识管理的新领域 [J]. 大学图书馆学报，2009, 27 (1): 30-37. (Qin Changjiang, Hou Hanqing. Mapping knowledge domain: A new field of information management and knowledge management [J]. Journal of Academic Libraries, 2009, 27 (1): 30-37.)

[21] 朱木易洁，鲍秉坤，徐常胜. 知识图谱发展与构建的研究进展 [J]. 南京信息工程大学学报, 2017, 9 (6): 575-582. (Zhu Muyijie, Bao Bingkun, Xu Changsheng. Research progress on development and construction of knowledge graph [J]. Journal of Nanjing University of Information Science and Technology, 2017, 9 (6): 575-582.)

[22] Hirschman L, Sager N. Automatic information formatting of a medical sublanguage [J]. Sublanguage: Studies of Language In Restricted Semantic Domains, 1982.

[23] 郑小林，王维维，扈中凯，等. 一种基于深度学习的中文医学知识图谱构建方法：中国，G06F17//30 [P]. 2017-05-31. (Zheng Xiaolin, Wang Weiwei, Hu Zhongkai, *et al.* A Chinese medical knowledge map construction method based on deep learning: China, G06F17//30 [P]. 2017-05-31.)